發表於華語文能力測驗成果發表會，國立台灣師範大學，2011 年 6 月 25 日

「華語文口語能力測驗」評分者一致性探討

張可家　施泰亨　藍珮君

一、測驗簡介

華語文口語能力測驗(前稱為 Test Of Proficiency-Huayu Speaking，現英文名稱更名為 Test of Chinese as a Foreign Language-Speaking,簡稱 TOCFL- Speaking)，專為母語非華語之人士研發的一種外語/第二語言口語能力測驗，測驗形式為電腦化測驗，口語考試題目透過電腦螢幕與耳機播放，題目播放完畢後，給予考生準備時間,準備時間結束後,考生以麥克風回答問題,錄音檔案將記錄在電腦裡。基於華語學習者的實際口語需求，以「溝通任務」為導向。

等級規劃主要參照CEFR(The Common European Framework of Reference for Languages：Learning, Teaching, Assessment) 口語能力指標描述，同時亦參考ILR（Interagency Language Roundtable）口語能力指標描述、ACTFL中文口語能力指標等資料。測驗等級分為基礎級、進階級、高階級、流利級，分別對應到CEFR的A2、B1、B2、C1等級，於2008年研擬出華語文口語能力指標，表1為TOCFL-Speaking與CEFR口語能力等級對應及華語文口語能力指標。2008年至2011年止，基礎級和進階級已舉辦過10場預試，高階級也於今年舉辦了第一次全國預試，流利級則尚在研發階段。

表1 TOCFL-Speaking與CEFR能力等級對應及華語文口語能力指標

TOCFL Speaking	CEFR	華語文口語能力指標
基礎級	A2	1. 能簡單敘述日常生活中熟悉的事物。 2. 能根據表格、海報、圖片、短片等，簡單描述內容。
進階級	B1	1. 能簡單、連貫地表達個人熟悉或感興趣的話題。 2. 能簡單、有次序地敘述個人經驗、理想或夢想。 3. 能簡單、清楚地說明計畫或事件；必要時，能提出理由支持自己的看法。
高階級	B2	1. 對於感興趣的話題，能清晰、流利地表達意見。 2. 對於一般性議題或有爭議的內容，能提出個人見解、舉例，並有組織地詳細說明理由。

流利級	C1	1. 對於複雜主題或各類話題，能流利、詳細地講解、描述。 2. 能得體地應對反對意見，論證層次清晰，結論完整合理。

測驗設計上，基礎級題目著重經驗的描述以及對事物的喜好，並回答與日常生活有關的話題；進階級題目著重情感表達，並且能根據所提供的資料說出自己的意見和想法；高階級的題目著重陳述意見並提出理由支持自己的論點；流利級題型尚在研發階段。

基礎級測驗共有兩個部分，第一部分為熱身題，目的在使學生能熟悉測驗方式，共 3 題，無準備時間，回答時間 40 秒；第二部分題型包括經驗描述、圖片或影片描述、訊息說明三大題型，共 5 題；進階級測驗共有三個部分，第一部分為熱身題，目的在使學生能熟悉測驗方式，共 3 題，無準備時間，回答時間 40 秒；第二及第三部分題型包括經驗描述、訊息說明、陳述意見三大題型共 6 題；高階級題型尚在預試階段，流利級題型則正在研發中。基礎級、進階級題型見表 2。

表 2 華語文口語能力題型

等級 / 部分	基礎級	進階級
第一部分	熱身題	熱身題
第二部分	經驗描述	經驗描述
	圖片或影片描述	訊息說明
	訊息說明	
第三部分	----	陳述意見

計分方式，除了熱身題不計分以外，其餘題型皆採用0-5級分之評分方式，3級分為通過門檻，評分者依照內容組織、表達能力以及語言運用三大方向對考生回答內容進行評分，考生在每一題的回答均會得到一個分數。評分著重於考察考生能否在特定情境下，藉由口語表達有效地傳遞訊息。

二、主觀式評分

口語能力測驗是一種表現測驗(performance assessment)，考生的成績由評分者根據評分原則來進行評分，而評分涉及到評分者主觀的判斷，因此有必要對主觀式評分有相當的認識，主觀式評分方式主要分為兩大類：一為分析式評分，一為整體式評分。已經有許多學者討論過其優缺點（Bachman,1988; Bachman & Savignon, 1986; Douglas & Smith, 1997; Fulcher ; Ingram & Wylie, 1993; Underhill, 1987）。本會最後採用整體式為主，分析式為輔的評分形式建立評分原則。目的希望能保留這兩種評分方式的優點，盡量避免缺點。以下是融合這兩種評分方式

的優點簡述：（一）能評定考生整體的口語表達溝通能力（二）評分者容易依循（三）方便省時（四）可提供使用者更多資訊（五）避免評分者在評分時標準不一致。

選定了評分的形式之後，接下來就是進行評分原則的撰寫。Fulcher（2003：91, 92）提到口語測驗評分原則（rating scale）的建立主要有兩種方式：一為直觀式（Intuitive methods），一為實證式（Empirical methods）。本會依照Fulcher所提出的直觀式評分原則流程，研擬華語文口語能力評分原則。首先經由專家的判斷，先草擬出一套量表，接著邀請學者專家組成委員會進行初步修訂，最後進入實際使用的階段，經過實際的使用，逐步修正。以下是較詳細的說明：

- 專家的判斷（Expert judgment）－所謂的專家包含有經驗的老師或是語言測驗研發人員，藉由他們的專業判斷，依照該測驗的性質，參考其他測驗的評分原則、教學計畫或是需求分析等資訊，草擬出評分原則。
- 委員會（Committee）－邀請數位專家學者組成諮詢委員會，一起討論評分原則用字遣詞的適當性。
- 經驗累積（Experiential）－評分者經由實際使用評分原則之後，再逐步加以修改。此方式為目前最常見的評分原則建立的方式。

本會依據上述流程著手先草擬出評分原則，接著根據諮詢委員的建議做初步的修正，最後實際運用在預試評分工作當中，藉由與評分者不斷討論後再行修改。華語文口語能力測驗評分原則，見附錄一。

確立了評分方式和評分原則之後，評分員的訓練也是相當重要的一環，本會對評分員的訓練制訂了一套標準化流程：首先將評分教師依照等級分組訓練。每次評分工作會舉辦兩次評分會議，第一次主要是做標準化的工作，藉由評分教師試評，統一大家的評分標準，然後讓老師獨立進行正式評分工作，評分工作結束之後，再舉辦第二次評分會議，進行討論，一方面再次調整不一致的部分，一方面再次確定各級分樣本，詳細評分訓練流程見表3。

表 3 華語文口語能力測驗評分訓練標準化流程

階段	工作項目	內容
1	第一次評分會議 前置作業	口語研發人員從考生音檔中挑選範例音檔做為第一次評分會議的試評音檔。
2	第一次評分會議	邀請核心評分老師參與評分會議，現場進行試評工作，並依據試評結果面對面討論，建立評分共識。
3	評分	核心老師將音檔以及相關文件帶回，進行為期一個月的評分工作。

4	第二次評分會議 前置作業	老師們繳交評分結果以及意見表，再由口語研發人員加以整理，彙整出需要討論的音檔以及問題。
5	第二次評分會議	邀請核心評分老師面對面討論，針對評分結果不一致的音檔，確立共識。
6	評分結果分析	將評分結果交由統計人員作更進一步的分析，以得知核心評分老師彼此之間的一致性,還有自己本身評分的穩定性等重要資料。

三、評分者一致性

經過了一連串持續且長期的評分訓練後,我們進一步探討評分者評分一致性的變化情形，包含與其他評分者之間的一致性，以及評分者自身的一致性。研究方法為收集2009年2月至12月五次華語文口語能力測驗-基礎級預試評分結果，採用多面向 Rasch 測量模式(many-facet Rasch measurement)分析軟體 FACETS，探討評分者評分一致性的變化情形,從評分者嚴格度(rater severity)探討評分者之間評分一致性、並從評分者面向之適配度(fit)探討評分者自身評分一致性。研究目的在探討接受持續且長期的評分訓練後,評分者之間與評分者自身的評分一致性的變化情況。

表 4 至表 8 為五次評分訓練後評分者的一致性的估計結果,在嚴格度欄位，數值為正表示偏嚴格，負值表示偏寬鬆，數值越大表示給分越嚴格，越小則表示越寬鬆。五次評分訓練後評分者嚴格度的差異分別為 1.09 logits、0.93 logits、0.31 logits、0.32 logits 以及 0.31 logits，顯示經過幾次評分訓練後，評分者之間嚴格度的差異呈現逐漸縮小的趨勢。

適配度統計值是指 MFRM 模式中，觀察到的評分與預期評分結果的適配情形,Lunz、Wright 和 Linacre(1990)建議可接受的範圍為 0.6 至 1.5(引自 Engelhard, 1992)；Linacre(2002)建議用 0.5 為低標，1.5 為高標，也有其他研究建議使用比較嚴格的標準(0.7-1.3)(McNamara, 1996；Bond & Fox, 2001；引自 Eckes, 2005)。在評分者內給分一致性方面，可以發現，即使採用較嚴格的標準(0.7-1.3)進行審視，五次評分結果，所有評分者的 infit 數值均落在適配的範圍之內(0.94-1.07、0.84-1.10、0.83-1.25、0.91-1.19、0.81-1.18)，表示評分者內給分一致性良好，評分者參與評分訓練後，實際進行給分時能保持相當的穩定性，不會過於偏離自身的標準，詳細數值見表 4 至表 8。

研究結果顯示：1.經過多次且密集的評分訓練後，評分者之間嚴格度分佈落差逐漸縮小，但整體評分嚴格度達到顯著差異水準，顯示評分者之間的嚴格度仍有不同，仍無法完全消除評分者個別的差異，評分者間的一致性仍有待提升。2.評分者自身的給分一致性良好，適配度數值均在可接受的範圍內，顯示長期且持

續的評分訓練對於評分者自身給分的一致性是有幫助的。

上述分析結果與 Engelhard(1992)、Weigle(1998)、Bonk 與 Ockey(2003)、Park(2004)以及 Eckes(2005)的研究結果相同，支持評分訓練對於給分極端的評分者能改善其嚴格度，但不能完全降低評分嚴格度的差異，評分訓練無法使評分者間的嚴格度達到一致，但是有助於提升評分者內給分一致性的論點。

由於此次研究受限於參加的評分者人數較少，且評分教師有時因個人因素無法參與每次培訓，是較為可惜的地方。未來我們將增加評分者的人數，以期能更明顯看出評分訓練對於評分者嚴格度的影響，當人數增加時，嚴格度的落差是否仍然與先前相當，或是變得更大。另外，也會用多面向 Rasch 測量模式持續針對進階級評分者評分一致性進行探討分析。

表 4　2009/02 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.38	0.10	0.98	0.99
B	0.33	0.10	0.94	0.92
A	-0.71	0.10	1.07	1.06

RMSE 0.10　Adj S.D. 0.61　Separation 6.08　Reliability 0.97
Fixed (all same) chi-sqaure: 76.8　d.f.: 2　sig: 0.00

表 5　2009/05 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.46	0.12	0.92	0.90
B	0.33	0.12	1.10	1.06
A	0.12	0.12	0.84	0.84
E	-0.44	0.12	1.07	1.07
D	-0.47	0.12	1.09	1.09

RMSE 0.13　Adj S.D. 0.41　Separation 3.30　Reliability 0.92
Fixed (all same) chi-sqaure: 47.2　d.f.: 4　sig: 0.00

表 6　2009/07 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
B	0.22	0.10	0.83	0.87
A	-0.04	0.10	0.86	0.84
D	-0.09	0.09	0.94	0.96
E	-0.09	0.09	1.25	1.30

RMSE 0.10　Adj S.D. 0.11　Separation 1.15　Reliability 0.57
Fixed (all same) chi-sqaure: 6.8　d.f.: 3　sig: 0.08

表 7　2009/10 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.18	0.08	1.19	1.17
C	0.05	0.08	0.92	0.92
A	-0.09	0.08	0.96	0.95
B	-0.14	0.08	0.91	0.91

RMSE 0.08　Adj S.D. 0.12　Separation 1.57　Reliability 0.71
Fixed (all same) chi-sqaure: 10.4　d.f.: 3　sig: 0.02

表 8　2009/12 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.14	0.08	1.14	1.14
B	0.03	0.08	0.82	0.82
A	0.00	0.09	0.81	0.81
C	-0.17	0.08	1.18	1.18

RMSE 0.08　Adj S.D. 0.10　Separation 1.28　Reliability 0.62
Fixed (all same) chi-sqaure: 8.6　d.f.: 3　sig: 0.04

附錄一

基礎級評分原則

級分	內容組織	表達能力	語言運用
5	內容與題目要求相關，內容豐富，話語多有組織。	語速適中，偶有停頓；詞語重複次數少；語音大致清楚，偶有偏誤，聽者都能理解。	已能掌握基本詞彙和語法，仍有少許偏誤。
4	回答與題目要求相關，內容尚稱豐富，話語尚有組織。	語速適中，常有停頓；詞語重複次數少；語音尚稱清楚，偶有偏誤，聽者幾乎都能理解。	已能大致掌握基本詞彙和語法，仍偶有偏誤。
3	內容與題目要求相關，內容稍嫌不足， 話題無法擴展 ，組織較差。	語速緩慢，常有停頓；詞語重複次數多；部份語音不正確，聽者尚能理解。	詞彙量仍有限，使用大致適當，尚能掌握基本語法結構，仍常有偏誤。
2	內容與題目要求相關但不完整，內容不足，組織甚差。	語速過於緩慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解。	詞彙量有限，且多不適當 ，尚未掌握基本語法結構。
1	內容與題目要求不相關。	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解。	詞彙量極有限，且不適當，無基本語法結構。
0	考生靜默，沒回答		

進階級第二部分評分原則

級分	內容組織	表達能力	語言運用
5	內容與題目要求相關，內容豐富，話語有組織	語速適中，偶有停頓；詞語重複次數甚少；語音清楚，較少偏誤，聽者都能理解	具備足夠的詞彙量和語法，能適當地使用，仍有少許偏誤
4	內容與題目要求相關，內容尚稱豐富，話語多有組織	語速適中，偶有停頓；詞語重複次數少；語音大致清楚，偶有偏誤，聽者幾乎都能理解	具備足夠的詞彙量和語法，能大致適當使用，偶有偏誤
3	內容與題目要求相關，但豐富性尚不足，部份話語缺乏組織	語速稍慢，常有停頓；詞語重複次數多；語音尚稱清楚，不時有偏誤，聽者尚能理解	尚有足夠的詞彙量和語法，尚能適當使用，不時有偏誤
2	內容與題目要求相關但不完整，內容不足，話題無法擴展，組織較差	語速過慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解	詞彙量有限，且多不適當，僅能掌握基本語法結構
1	內容與題目要求不相關	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解	詞彙量極有限，且不適當，僅能掌握部份基本語法結構
0	考生靜默，沒回答		

進階級第三部分評分原則

級分	內容組織	表達能力	語言運用
5	說出自己的選擇或意見，並能條理清楚，有步驟、有層次的說明解釋其理由，以支持自己的論點，話語有組織	語速適中，偶有停頓；詞語重複次數甚少；語音清楚，較少偏誤，聽者都能理解	具備足夠的詞彙量和語法，能適當地使用，仍有少許偏誤
4	說出自己的選擇或意見，並能提出多個理由支持自己的論點，偶有重複說明的情況，但已能進一步解釋，話語多有組織	語速適中，偶有停頓；詞語重複次數少；語音大致清楚，偶有偏誤，聽者幾乎都能理解	具備足夠的詞彙量和語法，能大致適當使用，偶有偏誤
3	說出自己的選擇或意見，並能提出至少一個理由支持自己的論點，常有重複說明的情況，以致於無法進一步加以解釋，部份話語缺乏組織	語速稍慢，常有停頓；詞語重複次數多；語音尚稱清楚，不時有偏誤，聽者尚能理解	尚有足夠的詞彙量和語法，尚能適當使用，不時有偏誤
2	說出自己的選擇或意見，但無法清楚地提出理由支持自己的論點；話語多不連貫，組織較差	語速過慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解	詞彙量有限，且多不適當，僅能掌握基本語法結構
1	沒說出自己的選擇或意見	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解	詞彙量極有限，且不適當，僅能掌握部份基本語法結構
0	考生靜默，沒回答		

參考文獻

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A Critique of the ACTFL Oral Interview. *The Modern Language Journal,* 70(4), 380-390.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition,* 10(2): 149-64

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion. *Language Testing, 20*(1), 89-110.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series No. 9). Princeton, NJ: ETS.

Engelhard, G. (1992). The measurement of writing ability with a Many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Fulcher, G. (2003). *Testing Second Language Speaking.* Pearson Education.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4*(1), 1-21.

Underhill, N. (1987). *Testing Spoken Language.* Cambridge: Cambridge University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.